Document Analysis Research in the Year 2021

Daniel Lopresti, Bart Lamiroy

Document Analysis Research in the Year 2021*

Daniel Lopresti[1] and Bart Lamiroy[2]

[1] Computer Science and Engineering, Lehigh University,
Bethlehem, PA 18015, USA – lopresti@cse.lehigh.edu
[2] Nancy Université, INPL – LORIA, Campus Scientifique,
BP 239, 54506 Vandoeuvre Cedex, France – Bart.Lamiroy@loria.fr

Abstract. Despite tremendous advances in computer software and hardware, certain key aspects of experimental research in document analysis, and pattern recognition in general, have not changed much over the past 50 years. This paper describes a vision of the future where community-created and managed resources make possible fundamental changes in the way science is conducted in such fields. We also discuss current developments that are helping to lead us in this direction.

1 Introduction: Setting the Stage

The field of document analysis research has had a long, rich history. Still, despite decades of advancement in computer software and hardware, not much has changed in how we conduct our experimental science, as emphasized in George Nagy's superb keynote retrospective at the DAS 2010 workshop [11].

In this paper, we present a vision for the future of experimental document analysis research. Here the availability of "cloud" resources consisting of data, algorithms, interpretations and full provenance, provides the foundation for a research paradigm that builds on collective intelligence (both machine and human) to instill new practices in a range of research areas. The reader should be aware that this paradigm is applicable to a much broader scope of machine perception and pattern recognition – we use document analysis as the topic area to illustrate the discussion as this is where our main research interests lie, and where we can legitimately back our claims. Currently under development, the platform we are building exploits important trends we see arising in a number of key areas, including the World Wide Web, database systems, and social and collaborative media.

The first part of this paper presents our view of this future as a fictional, yet realizable, "story" outlining what we believe to be a compelling view of community-created and managed resources that will fundamentally change the way we do research. In the second part of the paper, we then turn to a more technical discussion of the status of our current platform and developments in this direction.

* This work is a collaborative effort hosted by the Computer Science and Engineering Department at Lehigh University and funded by a Congressional appropriation administered through DARPA IPTO via Raytheon BBN Technologies.

2 Document Analysis Research: A Vision of the Future

Sometime in the year 2021, Jane, a young researcher just getting started in the field, decides to turn her attention to a specific task in document analysis: given a page image, identify regions that contain handwritten notations.[3] Her intention is to develop a fully general method that should be able to take any page as input, although there is the implicit assumption that the handwriting, if present, covers only a relatively small portion of the page and the majority of the content is pre-printed text or graphics.

Note that in this version of the future, there is no such thing as "ground truth" – that term is no longer used. Rather, we talk about the *intent* of the author, the *product* of the author (*e.g.*, the physical page) [4], and the *interpretation* arrived at by a reader of the document (human or algorithm). There are no "right" or "wrong" answers – interpretations may naturally differ – although for some applications, we expect that users who are fluent in the language and the domain of the document will agree nearly all of the time.

The goal of document analysis researchers is to develop new methods that mimic, as much as possible, what a careful human expert would do when confronted with the same input, or at least to come closer than existing algorithms. Of course, some people are more "careful" or more "expert" than others when performing certain tasks. The notion of *reputation*, originally conceived in the early days of social networking, figures prominently in determining whose interpretations Jane will choose to use as the target when developing her new method. Members of the international research community – as well as algorithms – have always had informal reputations, even in the early days of the field. What is different today, in 2021, is that reputation has been formalized and is directly associated with interpretations that we use to judge the effectiveness of our algorithms, so that systems to support experimental research can take advantage of this valuable information. Users, algorithms, and even individual data items all have reputations that are automatically managed and updated by the system.

After Jane has determined the nature of the task, she turns to a well-known resource – a web server we shall call *DARE* (for "Document Analysis Research Engine") – to request a set of sample documents which she will use in developing and refining her algorithm. This server, which lives in the "cloud" and is not a single machine, has become the de facto standard in the field, just as certain datasets were once considered a standard in the past. There are significant differences, however. In the early days, datasets were simply collections of page images along with a single interpretation for each item (which was called the "ground-truth" back then). In 2021, the DARE server supports a fundamentally different paradigm for doing experimental science in document image analysis.

[3] Experience has taught us that we tend to be overly optimistic when we assume problems like this will be completely solved in the near future and we will have moved on to harder questions. Since we need a starting point for our story, we ask the reader to suspend skepticism on what is likely a minor quibble. Jane's problem can be replaced with any one that serves the purpose.

Jane queries the server to give her 1,000 random documents from the various collections it knows about. Through the query interface, she specifies that:

- Pages should be presented as a 300 dpi bitonal TIF image.
- Pages should be predominately printed material: text, line art, photographs, etc. This implies that the page regions have been classified somehow: perhaps by a human interpreter, or by another algorithm, or some combination. Jane indicates that she wants the classifications to have come from only the most "trustworthy" sources, as determined through publication record, citations to past work, contributions to the DARE server, *etc.*
- A reasonable number of pages in the set should contain at least one handwritten annotation. Jane requests that the server provide a set of documents falling within a given range.
- The handwritten annotations on the pages should be delimited in a way that is consistent with the intended output from Jane's algorithm. Jane is allowed to specify the requirements she would like the interpretations to satisfy, as well as the level of trustworthiness required of their source.

By now, the status of the DARE web server as the de facto standard for the community has caused most researchers to use compatible file formats for recording interpretations. Although there is no requirement to do so, it is just easier this way since so much data is now delivered to users from the server and no longer lives locally on their own machines. Rather than fight the system, people cooperate without having to be coerced.

The DARE server not only returns the set of 1,000 random pages along with their associated interpretations, it also makes a permanent record of her query and provides a URL that will return exactly same set of documents each time it is run. Any user who has possession of the URL can see the parameter settings Jane used. The server logs all accesses to its collections so that members of the research community can see the history for every page delivered by the server.

In the early days of document image analysis research, one of the major hurdles in creating and distributing datasets were the copyright concerns. In 2021, however, the quantity of image-based data available on the web is astounding. Digital libraries, both commercially motivated and non-profit, present billions of pages that have already been scanned and placed online. While simple OCR results and manual transcriptions allow for efficient keyword-based searching, opportunities remain for a vast range of more sophisticated analysis and retrieval techniques. Hence, contributing a new dataset to the DARE server is not a matter of scanning the pages and confronting the copyright issues one's self but, rather, the vast majority of new datasets are references (links) to collections of page images that already exist somewhere online. Access – whether free or through subscription services – is handled as though well-developed mechanisms (including user authentication, if it is needed) that are part of the much bigger web environment.

With dataset in hand, Jane proceeds to work on her new algorithm for detecting handwritten annotations. This part of the process is no different from the way researchers worked in the past. Jane may examine the pages in the dataset

she was given by the DARE server. She uses some pages as a training set and others as her own "test" set, although this is just for development purposes and never for publication (since, of course, she cannot prove that the design of her algorithm was not biased by knowing what was contained in this set).

While working with the data, Jane notices a few problems. One of the page images was delivered to her upside down (rotated by 180 degrees). These sorts of errors, while rare, arise from time to time given the enormous size of the collections on the DARE server. In another case, the TIF file for a page was unreadable, at least by the version of the library Jane is using. Being a responsible member of the research community (and wanting her online reputation to reflect this), Jane logs onto the DARE server and, with a few mouse clicks, reports both problems – it just takes a minute. Everyone in the community works together to build and maintain the collections delivered via the web server. Jane's bug reports will be checked by other members of the community (whose reputations will likewise rise) and the problem images will be fixed in time.

In a few other cases, Jane disagrees with the interpretation that is provided for the page in question. In her opinion, the bounding polygons are drawn improperly and, on one page, there is an annotation that has been missed. Rather than just make changes locally to her own private copies of the annotation files (as would have happened in the past), Jane records her own interpretations on the DARE server and then refreshes her copies. No one has to agree with her, of course – the previous versions are still present on the server. But by adding her own interpretations, the entire collection is enriched. (At the same time Jane is doing her own work, dozens of other researchers are using the system.) The DARE server provides a wiki-like interface with text editing and graphical markup tools that run in any web browser. Unlike a traditional wiki, however, the different interpretations are maintained in parallel. The whole process is quite easy and natural. Once again, Jane's online reputation benefits when she contributes annotations that other users agree with and find helpful.

After Jane is done fine-tuning her algorithm, she prepares to write a paper for submission to a major conference. This will involve testing her claim that her technique will work for arbitrary real-world pages, not just for the data she has been using (and becoming very familiar with) for the past six months. She has two options for performing random, unbiased testing of her method, both of which turn back to the DARE server.[4] These are:

Option 1: Jane can "wrap" her code in a web-service framework provided by the DARE server. The code continues to run on Jane's machine, with the DARE server delivering a random page image that her algorithm has not seen before, but that satisfies certain properties she has specified in advance. Jane's algorithm performs its computations and returns its results to the DARE server within a few seconds. As the results are returned to the DARE server, they are compared to existing interpretations for the page in question.

[4] All top conferences and journals now require the sort of testing we describe here. This is a decision reached through the consensus of the research community, not dictated by some authority.

These could be human interpretations or the outputs from other algorithms that have been run previously on the same page.

Option 2: If she wishes, Jane can choose to upload her code to the DARE server, thereby contributing it to the community and raising her reputation. In this case, the server will run her algorithm locally on a variety of previously unseen test pages according to her specifications. It will also maintain her code in the system and use it in future comparisons when other researchers test their own new algorithms on the same task.

At the end of the evaluation, Jane is provided with:

- A set of summary results showing how well her algorithm matched human performance on the task.
- Another set of summary results showing how well her algorithm fared in comparison to other methods tested on the same pages.
- A *certificate* (*i.e.*, a unique URL) that guarantees the integrity of the results and which can be cited in the paper she is writing. Anyone who enters the certificate into a web browser can see the summary results of Jane's experiment delivered directly from the (trusted) DARE web server, so there can be no doubt what she reported in her paper is true and reproducible.

When Jane writes her paper, the automated analysis performed by the DARE server allows her to quantify her algorithms performance relative to that of a human, as well as to techniques that were previously registered on the system. Of course, given the specifics of our paradigm, performances can only be expressed in terms of statistical agreement and, perhaps, reputation, but perhaps not in terms of an absolute ranking of one algorithm with respect to another. Ranking and classification of algorithms and the data they were evaluated on will necessarily take more subtle and multi-valued forms. One may argue that having randomly selected evaluation documents for certification can be considered as marginally fair, since there is a factor of chance involved. While this is, in essence, true, the fact that the randomly generated dataset is available for reproduction (*i.e.* once generated, the certificate provides a link to the exact dataset used to certify the results), anyone arguing that the result was obtained on an unusually biased selection can access the very same data and use it in evaluating other algorithms.

It was perhaps a bit ambitious of Jane to believe that her method would handle all possible inputs and, in fact, she learns that her code crashes on two of the test pages. The DARE server allows Jane to download these pages to see what is wrong (it turns out that she failed to dimension a certain array to be big enough). If Jane is requesting a certificate, the DARE server will guarantee that her code never sees the same page twice. If she is not requesting a certificate, then this restriction does not apply and the server will be happy to deliver the same page as often as she wishes.

Unlike past researchers who had the ability to remove troublesome inputs from their test sets in advance, the DARE server prohibits such behavior. As a result, it is not uncommon for a paper's authors to report, with refreshing

honesty, that their implementation of an algorithm matched the human interpretation 93% of the time, failed to match the human 5% of the time, and did not complete (*i.e.*, crashed) 2% of the time.

When other researchers read Jane's paper, they can use the URL she has published to retrieve exactly the same set of pages from the DARE server.[5] If they wish to perform an unbiased test of their own competing method, comparing it directly to Jane's – and receive a DARE certificate guaranteeing the integrity of their results – they must abide by the same rules she did.

In this future world, there is broad agreement that the new paradigm introduced (and enforced) by the DARE server has improved the quality of research. Results are now verifiable and reproducible. Beginning researchers no longer waste their time developing methods that are inferior to already-known techniques (since the DARE server will immediately tell you if another algorithm did a better job on the test set you were given). The natural (often innocent) tendency to bias an algorithm based on knowing the details of the test set have been eliminated. The overuse of relatively small "standard" collections that was so prevalent in the early days of the field is now no longer a problem.

The DARE server is not foolproof, of course – it provides many features to encourage and support good science, but it cannot completely eliminate the possibility of a malicious individual abusing the system. However, due to its community nature, all records are open and visible to every user of the system, which increases the risk of being discovered to the degree that legitimate researchers would never be willing to take that chance.

Looking back with appreciation at how this leap forward was accomplished, Jane realizes that it was not the result of a particular research project or any single individual. Rather, it was the collective effort and dedication of the entire document analysis research community.

3 Script and Screenplay for the Scenario

The scenario just presented raises a number of fundamental questions that must be addressed before document analysis research can realize its benefits. In this section we develop these questions and analyze the extent to which they already have (partial or complete) answers in the current state-of-the-art, those which are open but that can be answered with a reasonable amount of effort, and those which will require significant attention by the community before they are solved.

We also refer to a proof-of-concept prototype platform for Document Analysis and Exploitation (DAE – not to be confused with DARE), accessible at `http://dae.cse.lehigh.edu`, which is capable of storing data, meta-data and interpretations, interaction software, and complete provenance as more fully described elsewhere [8, 9]. DAE is an important step in the direction of DARE, but still short of the grand vision described earlier.

[5] This form of access-via-URL is not limited to randomly generated datasets. Legacy datasets from the past are also available this way.

3.1 Main Requirements

What the scenario describes, in essence, is the availability of a well identified, commonly available resource that offers storage and retrieval of document analysis data, complex querying of this data, collective yet personalized markup, representation, evaluation, organization and projection of data, as well as archival knowledge of uses and transformations of data, including certified interactions.

Rather than offer monolithic chunks of data and meta-data or interpretations as in the case of current standard datasets, the envisioned resource treats data on a finer-grained level. This level of detail is illustrated in the scenario by the complexity of the queries the system should be capable of answering.

This data need not conform to a predefined format, but can be polymorphic, originating both from individual initiatives as well as from collective contributions. Regardless of how it is stored, it can also be retrieved and re-projected into any format. Annotations need not be human-contributed but can be the result of complete document analysis pipelines and algorithms. As a result, the resource can hold apparently contradictory interpretations of identical documents, when these stem from different transformation and analysis processes [8].

Formats and representations cannot be rigidly dictated in advance for our scenario to have a chance of succeeding. Past experience has shown that attempts to "coerce" a community into using a single set of conventions does not work; at best, it contributes to locking it into a limited subset of possible uses, stifling creativity. This is contradictory to the standpoint we have taken with respect to *ground-truth* (or rather lack thereof [6, 10, 14, 2]) and our preferring the term *interpretation*. This clearly advocates for as open as possible ways of representing data, keeping in mind, however, that abandoning any kind of imposed structure may make it impossible to realize our vision.

Turning now to the current DAE server, formats and representations are transparently handled by the system, since the user can define any format, naming, or association convention within our system. Data can be associated with image regions, image regions can be of any shape and format, there is no restriction on uniqueness or redundancy, so multiple interpretations are naturally supported. Because of its underlying data model and architecture, everything is queryable via SQL. The standard datasets that can be downloaded from the platform are no longer monolithic file collections, but potentially complex queries that generate these datasets on-the-fly [9].

Interactions with data are integrated in the DAE data model on the one hand (it represents algorithms as well as their inputs and outputs), but the model goes further by giving access to user-provided programs that can be executed on the stored data, thus producing new meta-data and interpretations. Queries like finding all results produced by a specified algorithm, class of algorithms, or user, can be used as an interpretation of a document, and can also serve as a benchmarking element for comparison with competitors. Since everything is hosted in the same context, it becomes possible to "certify" evaluations performed on the system.

3.2 Scientific (and Other) Challenges

While the DAE platform is a promising first step toward to the DARE paradigm, it still falls short in addressing some of the key concepts of the scenario we depicted in Section 2.

- Reputation is one suggestion we advanced to to treat disagreements between multiple interpretations. Not only can multiple interpretations arise from using the same data in different contexts, but there can be debate even in identical contexts when the data is "noisy." When such controversy arises, which would be the preferred interpretation to use? Here the notion of on-line reputation as practiced in Web 2.0 recommender systems may hold the key [12, 13]. Researchers and algorithms already have informal reputations within the community. Extending this to interpretations can provide a mechanism for deciding which annotations to trust.
  How this actually needs to be implemented and made user-friendly is an interesting question. Success in doing so would likely also solve the problem of deliberately malicious interpretations.
- Semantic clutter is another major issue that will inevitably occur and that has no straightforward solution under the current state-of-the-art. Semantic clutter arises when different contributors are unaware of each other's interpretative contexts and label data with either identical labels even though their contexts are completely different, or, conversely, with different labels although they share the same interpretation context. In the case of Jane, for instance, some image region may be labeled as *handwriting* but is, in fact, the interpretation of a printed word spelled that way, rather than denoting actual handwritten text. On the other hand, she might miss regions that have been labeled by synonyms such as *manuscript*, *hand annotated*, *writing* ... There are probably very good opportunities to leverage work on *folksonomies* [7, 3] to solve some of these issues, although our context is slightly different than the one usually studied in that community. Since, in our case, all data provenance is recorded and accessible, one can easily retrieve data produced by an algorithm using specific runtime parameters. This guarantees that the obtained data share the same interpretation context, and thus has some significant overlapping semantic value. Furthermore, since users are free to use and define their own, personal annotation tags, and since, again, provenance is recorded, one can assume that the semantics of a given user's tags will only evolve slowly over time, if at all. Exploring advanced formal learning techniques might yield a key to reducing the semantic cluttering mentioned earlier, or at least provide tools for users to discover interpretations that are likely to be compatible with their own context.
- Query Interfaces, and especially those capable of handling the complex expressions used in this paper, are still open research domains [1]. Their application is also directly related to the previously mentioned semantic issues. Semantic Web [5] and ontology-folksonomy combinations [7, 3] are therefore also also probable actors in our scenario. To make the querying really semantic in an as automated a way as possible, and by correctly capturing

the power of expressiveness as people contribute to the resource pool, the DAE platform will need to integrate adequate knowledge representations. This goes beyond the current storage of attributes and links between data. Since individual research contexts and problems usually require specific representations and concepts, contributions to the system will initially focus on their own formats. However, as the need for new *interpretations* arises, users will want to combine different representations of similar concepts to expand their experiment base. To allow them to do that, formal representations and semantic web tools will need to be developed. Although their seems to be intuitively obvious inter-domain synergies between all cited domains (*e.g.* data-mining query languages need application contexts and data to validate their models, while our document analysis targeted platform needs query languages to validate the scalability and generality of its underlying research) only widespread adoption of the paradigm described in this paper will reveal potentially intricate research interactions.

– Archiving and Temporal Consistency, concern a fundamentally crucial part of this research. Numerous problems arise relating to the comprehensive usage and storage of all the information mentioned in this paper. In short, and given that our scenario is built on a distributed architecture, how shall availability of data be handled? Since our concept relies on complete traceability and inter-dependence of data, annotations and algorithms, how can we guarantee long term conservation of all these resources when parts of them are third-party provided? Simple replication and redundancy may rapidly run into copyright, intellectual property, or trade secret issues. Even data that was initially considered public domain may suddenly turn out to be owned by someone and need to be removed. What about all derived annotations, interpretations, and results? What about reliability and availability from a purely operational point of view, if the global resource becomes so widely used that it becomes of vital importance?

4 Conclusion

In this paper, we have presented a vision for the future of experimental research in document analysis and described how our current DAE platform [8, 9] can exploit collective intelligence to instill new practices in the field. This forms a significant first step toward a crowd-sourced document resource platform that can contribute in many ways to more reproducible and sustainable machine perception research. Some of its features, such as its ability to host complex workflows, are currently being developed to support benchmarking contests.

We have no doubt that the paradigm we are proposing is largely feasible and we strongly believe that the future of experimental document analysis research will head in a new direction much like the one we are suggesting. When this compelling vision comes to pass, it will be through the combined efforts and ingenuity of the entire research community.

The DAE server prototype is open to community contributions. It is inherently cloud-ready and has the potential to evolve to support the grand vision of

DARE. Because of this new paradigm's significance to the international research community, we encourage discussion, extensions and amendments through a constantly evolving Wiki: `http://dae.cse.lehigh.edu/WIKI`. This Wiki also hosts a constantly updated chart of DAE platform features realizing the broader goals discussed in this paper.

References

1. Boulicaut, J.F., Masson, C.: Data mining query languages. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 655–664. Springer US (2010), 10.1007/978-0-387-09823-4_33
2. Clavelli, A., Karatzas, D., Lladós, J.: A framework for the assessment of text extraction algorithms on complex colour images. In: DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems. pp. 19–26. ACM, New York, NY, USA (2010)
3. Dotsika, F.: Uniting formal and informal descriptive power: Reconciling ontologies with folksonomies. International Journal of Information Management 29(5), 407–415 (October 2009)
4. Eco, U.: The limits of interpretation. Indiana University Press (1990)
5. Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., Stephens, S.: The semantic web in action. Scientific American December (2007)
6. Hu, J., Kashi, R., Lopresti, D., Nagy, G., Wilfong, G.: Why table ground-truthing is hard. In: 6th International Conference on Document Analysis and Recognition. pp. 129–133. IEEE Computer Society (2001)
7. Kim, H.L., Decker, S., Breslin, J.G.: Representing and sharing folksonomies with semantics. Journal of Information Science 36(1), 57–72 (February 2010)
8. Lamiroy, B., Lopresti, D.: A platform for storing, visualizing, and interpreting collections of noisy documents. In: Fourth Workshop on Analytics for Noisy Unstructured Text Data - AND'10. ACM International Conference Proceeding Series, IAPR, ACM, Toronto Canada (October 2010)
9. Lamiroy, B., Lopresti, D., Korth, H., Jeff, H.: How carefully designed open resource sharing can help and expand document analysis research. In: Agam, G., Viard-Gaudin, C. (eds.) Document Recognition and Retrieval XVIII. SPIE Proceedings, vol. 7874. SPIE, San Francisco, CA USA, (January 2011)
10. Lopresti, D., Nagy, G., Smith, E.B.: Document analysis issues in reading optical scan ballots. In: DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems. pp. 105–112. ACM, New York, NY, USA (2010)
11. Nagy, G.: Document systems analysis: Testing, testing, testing. In: Doerman, D., Govindaraju, V., Lopresti, D., Natarajan, P. (eds.) DAS 2010, Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems. p. 1 (2010), `http://cubs.buffalo.edu/DAS2010/GN_testing_DAS_10.pdf`
12. Raub, W., Weesie, J.: Reputation and efficiency in social interactions: An example of network effects. American Journal of Sociology 96(3), 626–654 (1990)
13. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review 24(1), 33–60 (2005)
14. Smith, E.H.B.: An analysis of binarization ground truthing. In: DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems. pp. 27–34. ACM, New York, NY, USA (2010)